IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Predicting Academic Course Preference Using Hadoop

Ms. Namrata Thakre[1], Mr. Hirendra Hajare[2]

M. Tech Student, Department of CSE, Ballarpur Institute of Technology (BIT), Ballarpur[1]

Assistant Professor, Department of CSE, Ballarpur Institute of Technology (BIT), Ballarpur[2]

Abstract: With the emergence of new technologies, new academic trends introduced into Educational system which results in large data which is unregulated and it is also challenge for students to prefer to those academic courses which are helpful in their industrial training and increases their career prospects. Another challenge is to convert the unregulated data into structured and meaningful information there is need of Data Mining Tools. Hadoop Distributed File System is used to hold large amount of data. The Files are stored in a redundant fashion across multiple machines which ensure their endurance to failure and parallel applications. Knowledge extracted using Map Reduce will be helpful indecision making for students to determine courses chosen for industrial trainings. In this research, we are deriving preferable courses for pursuing training for students based on course combinations. Here, using HDFS, tasks run over Map Reduce and output is obtained after aggregation of results.

Keywords: Distributed File System, data mining, educational data mining, Hadoop, MapReduce.

I. INTRODUCTION

Data mining is one of the most prominent areas in modern technologies for retrieving meaningful information from huge amount of unstructured and distributed data using parallel processing of data. There is huge advantage to Educational sector of following Data Mining Techniques to analyse data input from students, feedbacks, latest academic trends etc which helps in providing quality education and decision-making approach for students to increase their career prospects and right selection of courses for industrial trainings to fulfil the skill gap pertains between primary education and industry hiring students. Data Mining has great impact in academic systems where education is weighed as primary input for societal progress.

Big data is the emerging field of data mining. It is a term for datasets that are so large or complex that traditional data processing application software is incompetent to deal with them. Big data includes gathering of data for storage and analysis purpose which gain control over operations like searching, sharing, visualization of data, query processing, updating and maintain privacy of information. In Big data, here is extremely large dataset that is analysed computationally to reveal patterns, trends and associations. It deals with unstructured data which may include MS Office files, PDF, Text etc whereas structured data may be the relational data.

Hadoop is one technique of big data and answer to problems related to handling of unstructured and massive data. Hadoop is an open-source programming paradigm which performs parallel processing of applications on clusters. Big Data approach can help colleges, institutions, universities to get a comprehensive aspect about the students. It helps in answering questions related to the learning behaviours, better understanding and curriculum trends, and future course selection for students which helps to create captivating learning experiences for students. The problem of enormously large size of dataset can be solved using Map Reduce Techniques. Map Reduce jobs run over Hadoop Clusters by splitting the big data into small chunks and process the data by running it parallel on distributed clusters.

II. LITERATURE REVIEW

Jongwook Woo, "Apriori-Map/Reduce Algorithm." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.Apriori algorithm is implemented and high performance is achieved using Map Reduce Technique of Hadoop framework to collect item sets frequently occurred in dataset.

Xin YueYang,Zhen Liu,Yan Fu, **"MapReduce as a Programming Model for Association Rules Algorithm on Hadoop"**, Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on, pp. 99-102. IEEE, 2010. Author has mainly addressed the challenges of using Map Reduce model for computing parallel application of Apriori.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 6, June 2020

Big Data Techniques are the necessity in learning environments and present scenario with large amount of unstructured data and introduction of Massive open online courses in Education has stressed upon the need for data mining in Education.

B.Manjulatha, Ambica Venna, K.Soumya, "Implementation of Hadoop Operations for Big Data Processing in Educational Institutions", International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online) : 2320-9801, Vol. 4, Issue 4, April 2016. Tools of Data Mining like MangoDB, an open-sourcebdatabase and Apache Hadoop are discussed. Data Mining Techniques using these tools help students in choosing their course curriculum.

The term "big knowledge" is outlined as data that becomes thus giant that it cannot be processed exploitation standard strategies. the dimensions of the knowledge which might be thought-about to be huge knowledge could be a perpetually varied issue and newer tools area unit incessantly being developed to handle this huge data. so as to create sense out of this overwhelming quantity of knowledge it's typically Softened exploitation 5 V's: Velocity, Volume, Value, Variety, and Veracity.

Before MapReduce, doing this type of calculation would are troublesome. Currently programmers will handle this type issues with an ease. The advanced algorithms have been coded by the knowledge scientists for frameworks so that it becomes easy to use for the programmers. They don't want the department of PhD scientists to develop a whole complex framework. As MapReduce will work on network which provides a straightforward analysis. MapReduce is gaining much users as a result of the Apache Hadoop and Spark parallel computing systems. Let the programmers use MapReduce to run models over huge distributed sets of data and use advanced techniques of math and machine learning techniques so that we can predict the results easily realize patterns, uncover correlations, etc.

Proposed System

Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node. The input dataset is fed into the mapper and after passing through shuffle phase, reducer displays the output after aggregating the tuples obtained from mapper and are in the form of <key, value> pair.

III. METHODOLOGY

Dataset for Course Selection

Table I shows the list of course combinations taken by students for their industrial trainings

Table I. Dataset for Course Selection

Course Combination	Preferable Course
Java,J2EE	Android
HTML,Javascript	PHP
C,C++ Asp.Net	Asp.Net
J2EE,Java	Android
C,C++	Java

Here each row is considered as transaction, each comprising a combination of variables or item sets. The pattern of student's choice for industrial training course combinations is predicted after processing through MAP Reduce Hadoop Data Mining Technique shown in figure.

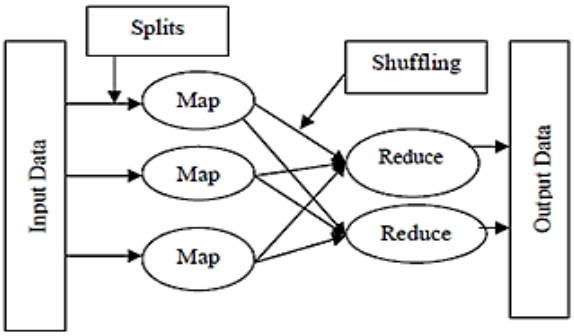

Fig 1: MapReduce Organization Chart

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 6, June 2020

The input dataset collected from students is shown in Table I is stored in the HDFS for MapReduce. The input data is then split into various clusters and provide it to the mapper that maps data to the output. The output from the mapper is represented as <key, value> pair. The output obtained from the mapper are then combined together in the combiner and then sent to the reducer.

Here, for organizing the work, Hadoop divides the task into Map and Reduce Tasks. The components of Hadoop Distributed File System are discussed below. The Map Reduce program transforms lists of input data elements into list of output data elements and it will be done using twice by Map and Reduce. The cluster running applications and name node information in the form of Web Interfaces using Hadoop.

The Organisation Structure of Map Reduce Framework is shown in Fig.1which represents that input data obtained from Table 1 first splits and then mapped followed with shuffling. The unstructured data after shuffling filtered to obtain output which is also called "Reduce Phase".

IV. SYSTEM DESIGN

System design is the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development "blends the perspective of marketing, design, and manufacturing into a single approach to product development," then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

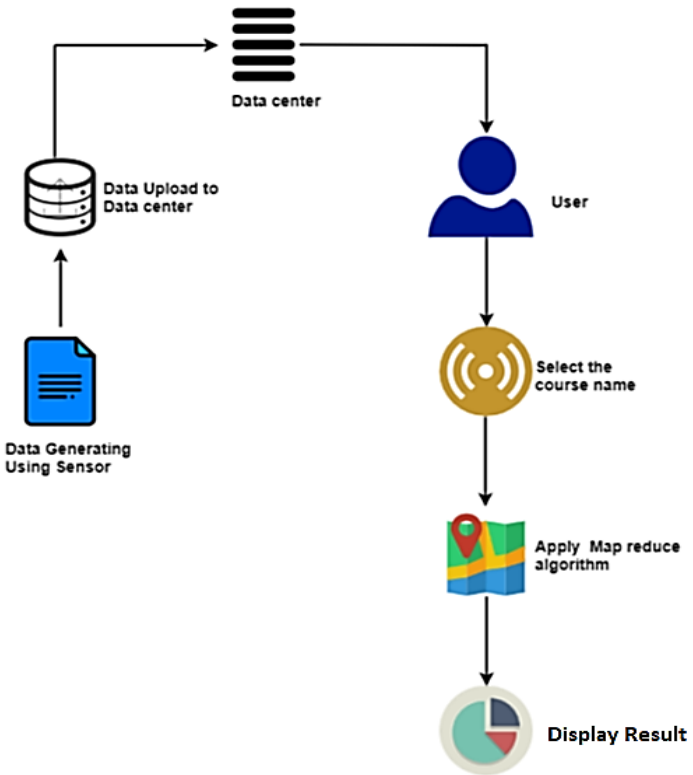

Fig: System Architecture

MapReduce has become of the foremost often used framework for processing of giant quantity of knowledge hold on in Hadoop cluster. It is used for multiprocessing of giant quantity of knowledge speedily. Firstly, it had been designed by google to produce the correspondence and cut back the fault tolerance of knowledge.

MapReduce processes the info within the type of key price pairs. we are able to select the key price pairs supported our alternative. We need to use the key price pairs for MapReduce as our schema isn't static. after we have static schema, we are able to use columns for analysing the info.

Map cut back API can furnish the next choices like process, multiprocessing of giant amounts of knowledge and high accessibility. The Map cut back work flow can undergoes totally different phases that stores the lead to HDFS with replications at the top. Job Trackers can do the work of checking all the Map cut back jobs that an acting on the Hadoop Cluster.

The Job huntsman can play an important role in planning jobs & it'll keeps the track of each map and cut back jobs. The task hunter can do the particular map and cut back jobs. Map cut back design principally consists of 2 process stages. 1st one is that the map stage & thus the opposite is cut back stage. Between these 2 stages there an extra stage referred to as intermediate stage that will the work of taking the input from the mappers and doing the tasks like shuffle, sort, mix etc.

V. CONCLUSION

The Map Reduce approach is used for running jobs over HDFS. Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node. The input dataset is fed into the mapper and after passing through shuffle phase, reducer displays the output after aggregating the tuples obtained from mapper and are in the form of <key, value> pair.

REFERENCES

[1]. Saptarshi Ray, “Big Data in Education”, Gravity, the Great Lakes Magazine, pp. 8-10, 2013.
[2]. Jongwook Woo, "Apriori-Map/Reduce Algorithm." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (World Comp), 2012.
[3]. Katrina Sin, Loganathan Muthu, "Application of Big Data in Education Data Mining and Learning Analytics – A Literature Review", ICTACT Journal on Soft Computing, ISSN: 2229-6956 (online), Vol 5, Issue 4, July 2015.
[4]. Shriram Raghunathan and Abtar Kaur, “Assessment of online interaction pattern using the Q-4R framework”, The International Lifelong Learning Conference, 2011.
[5]. B. Manjulatha, Ambica Venna, K.Soumya, "Implementation of Hadoop Operations for Big Data Processing in Educational Institutions", International Journal of Innovative Research in Computer & Communication Engineering, ISSN(Online) : 2320-9801,Vol. 4, Issue 4, April 2016.
[6]. N.Tajunisha, M.Anjali, "Predicting Student Performance Using MapReduce”, IJECS, Vol.4, Issue 1, Jan 2015, p. 9971-9976.
[7]. Shankar M.Patil, Praveen Kumar, “Data Mining Model for Effective Data Analysis of Higher Education Students Using MapReduce”, IJERMT,ISSN:2278-9359,Vol.6,Issue4,April2017.
[8]. Madhavi Vaidya, "Parallel Processing of cluster by Map Reduce", IJDPS, Vol.3, No.1,2012.
[9]. Harshawardhan S.Bhosale, Devendra P.Gadekar, "A Review Paper on Big Data and Hadoop",IJSRP,ISSN:2250-3153,Vol 4,Issue 10,Oct 2014.